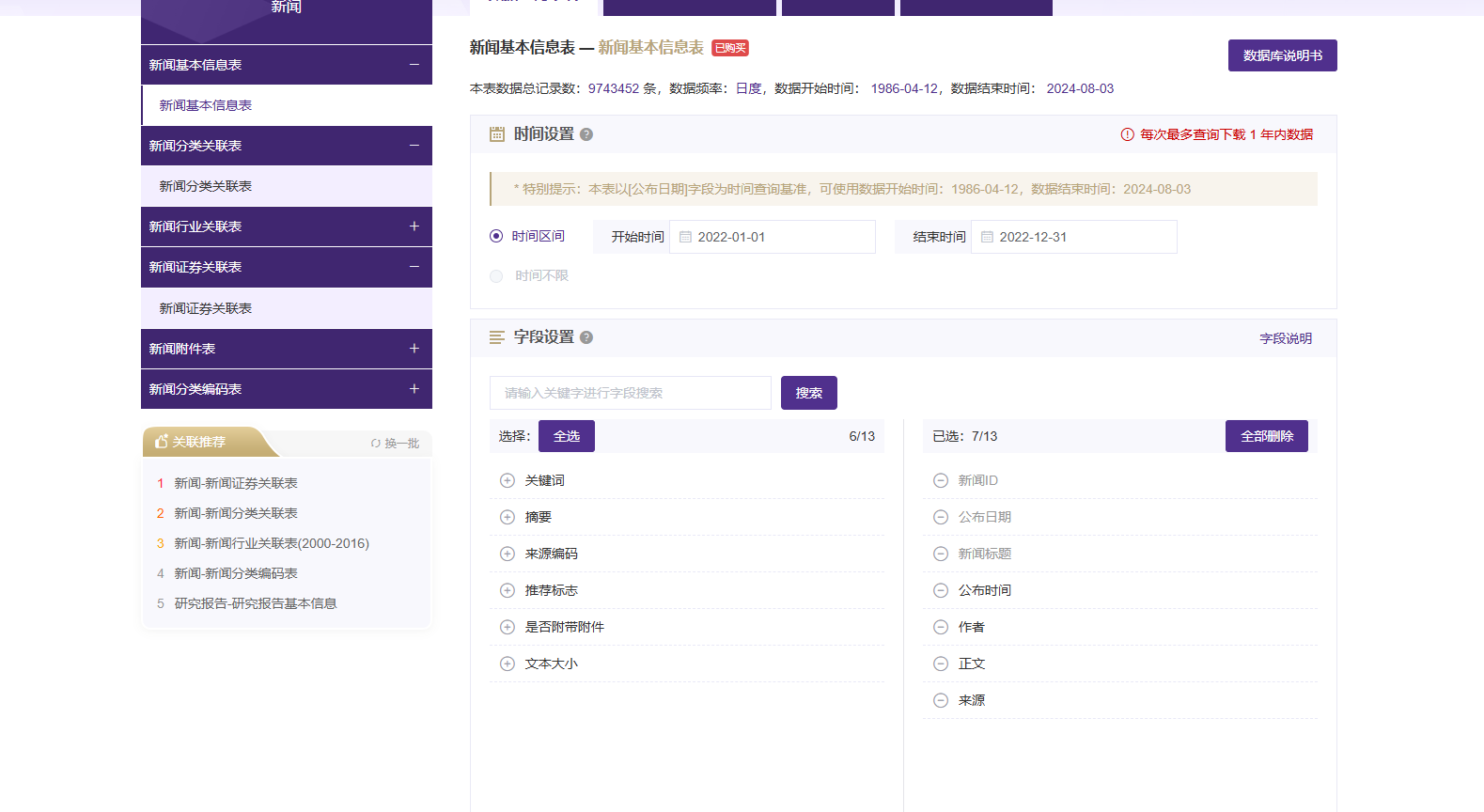
# 一、

## 1、

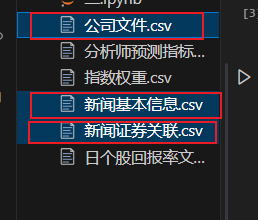
数据采集：



通过数据的时间筛选获取数据，选择下载csv格式数据。

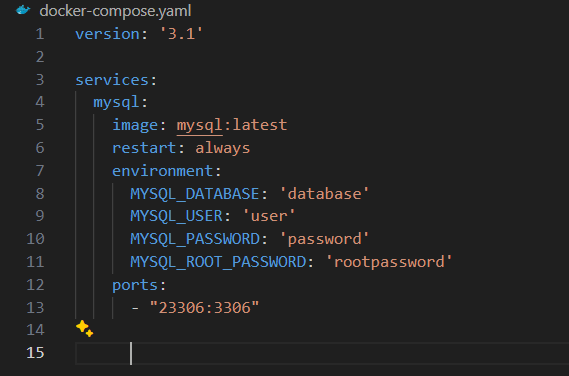
根据题目要求，需要下载的数据有 新闻基本信息表；新闻关联信息表；公司文件表。

完成下载后得到csv数据。



为将数据导入数据库，采用python读取数据，并使用mysql作为存储介质。

使用docker-compose创建数据库环境，相关参数如下：



由此搭建docker环境，并使用python的sqlalchemy库将数据转储到数据库中。

数据库介绍：

有哪些库 只使用一个数据库 database存储数据

有哪些表 按照问题1的要求，建立三张表，分别为新闻基本信息表、新闻证券关联表、公司文件信息表。

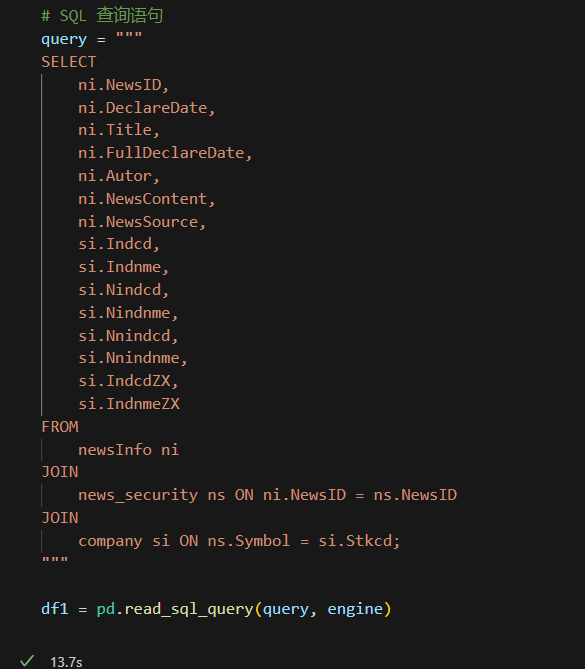
主键 新闻基本信息表采用newsid作为主键

新闻证券关联表采用newsid+symbol作为主键

公司文件信息表采用stkcd作为主键

表与表之间的联系 各表独立建立，不设置关联

每一部分如何实现的 问题要求结合各表数据获取指定数据，在完成建表后使用sql语句获取，结果如下：

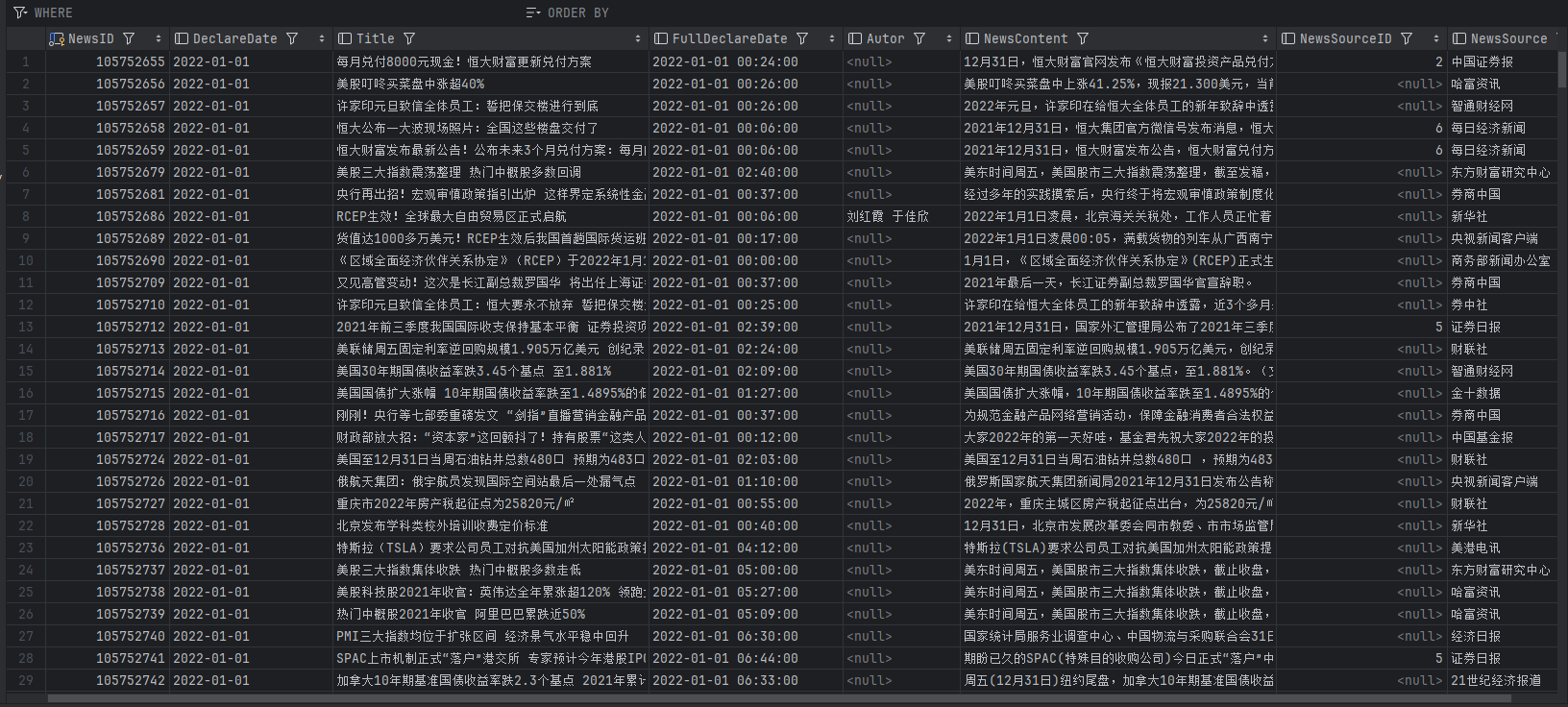


由此可以直接将数据缓存至程序中用于数据的进一步调用。

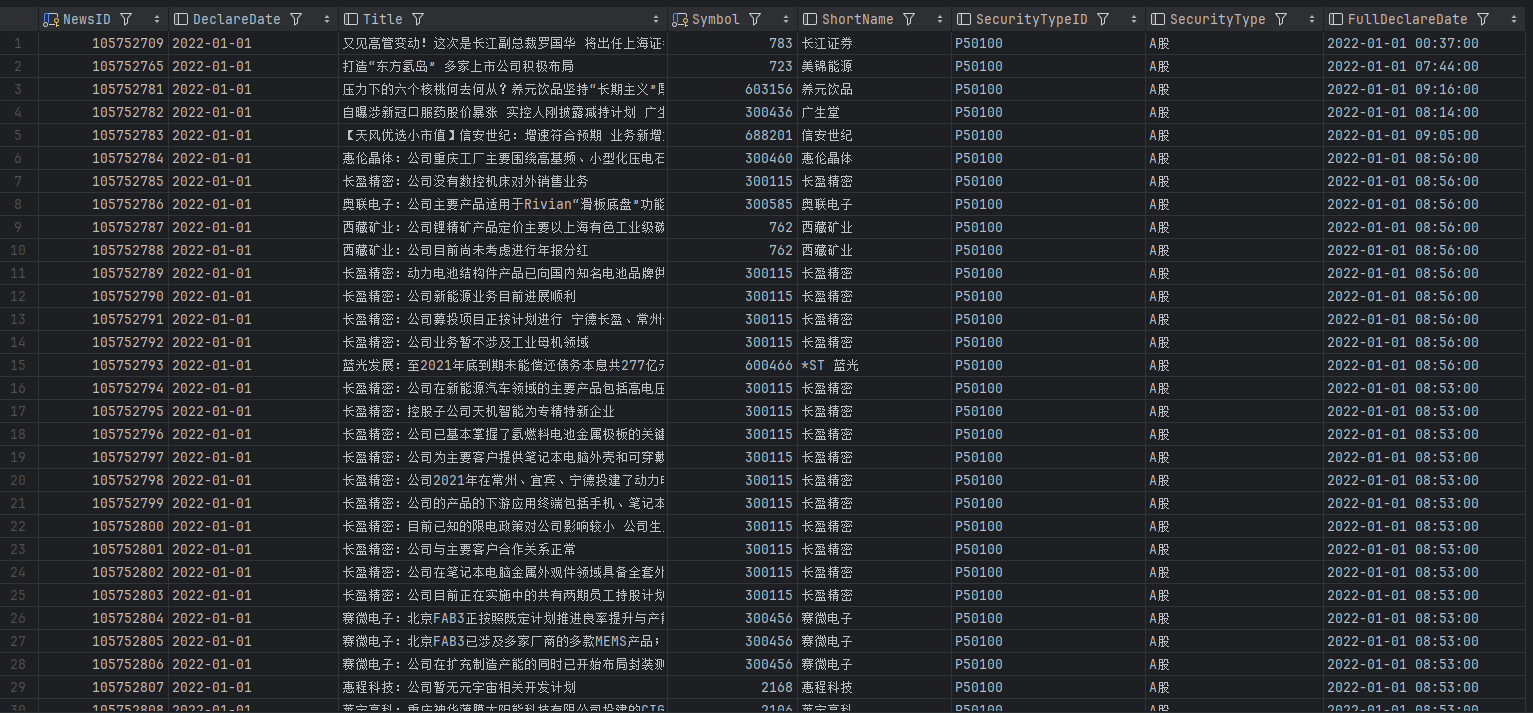
也可以直接使用create table在数据库中建立新表用于存储数据。

所建数据库截图（重要）+其他截图：

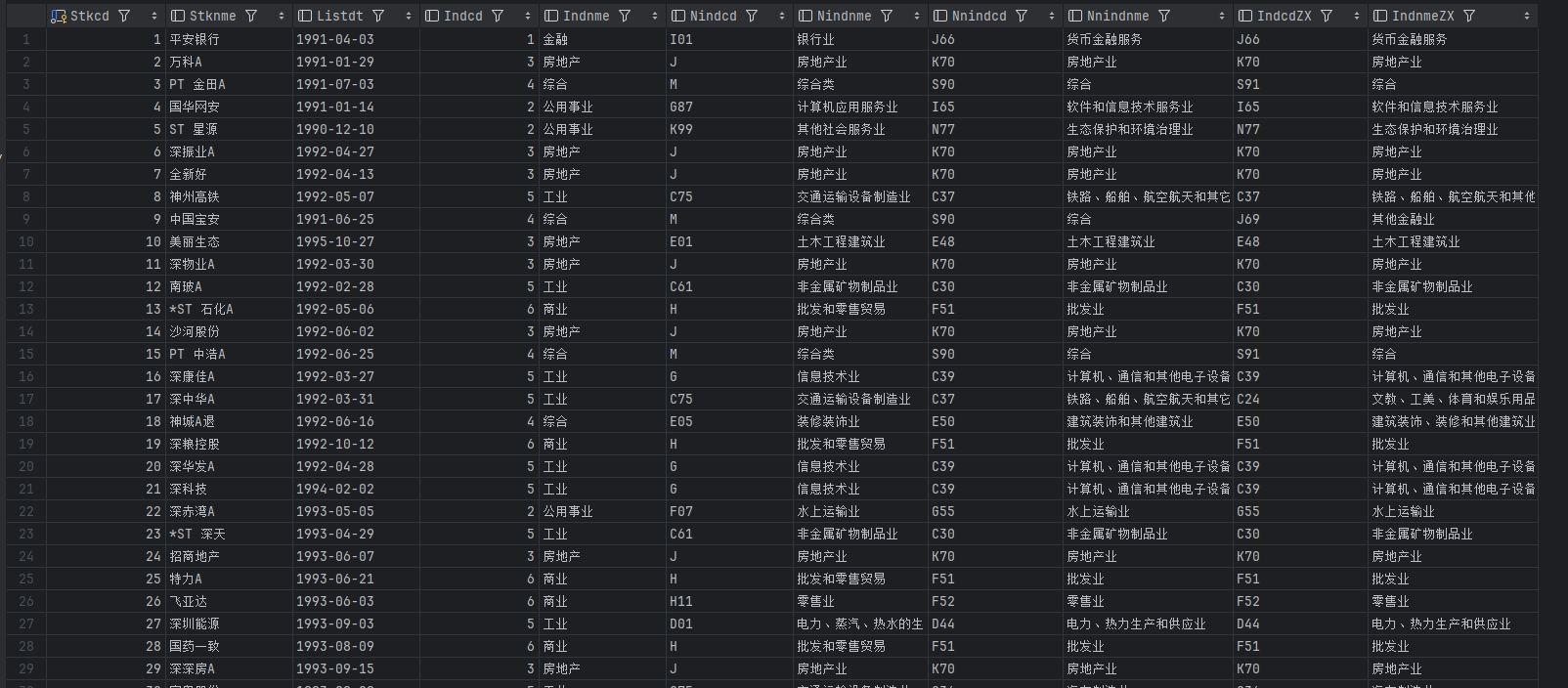
新闻基本信息表：



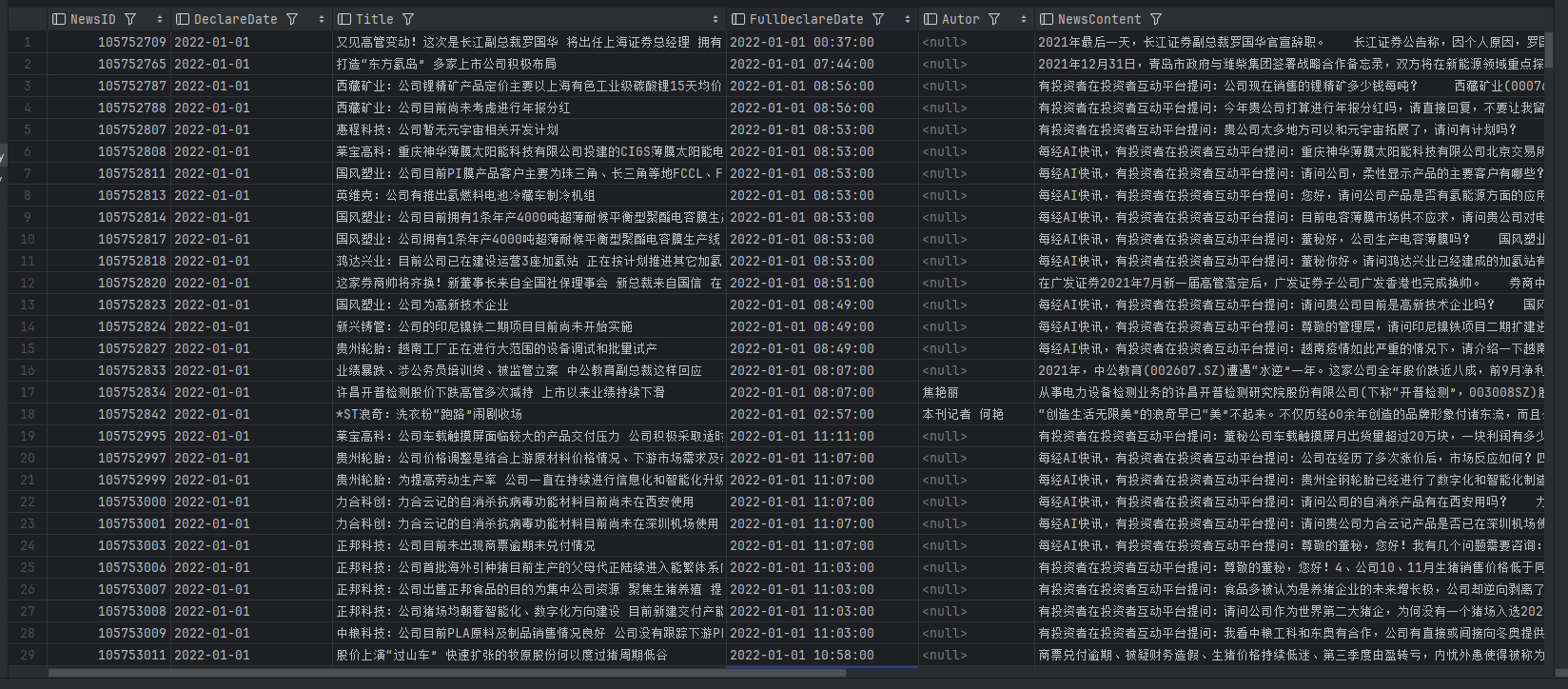
新闻证券关联表：

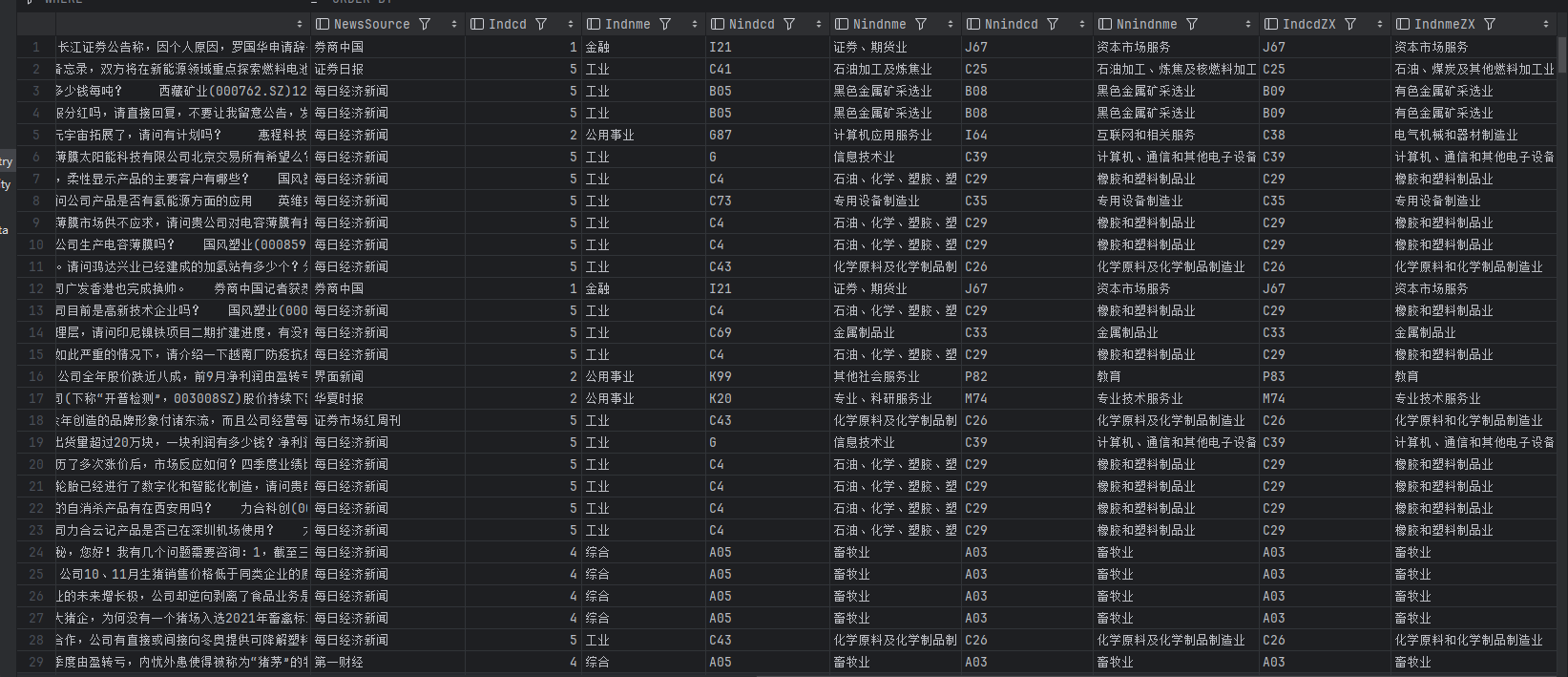


企业信息表：



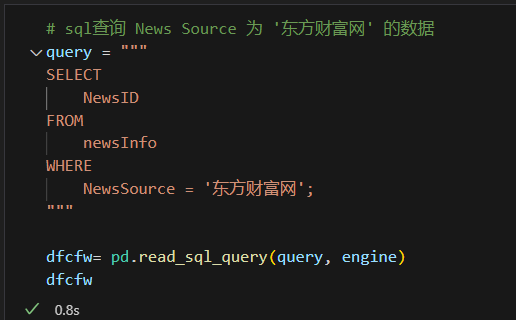
问题结果展示表：





（字段过多，采用两张图片展示）

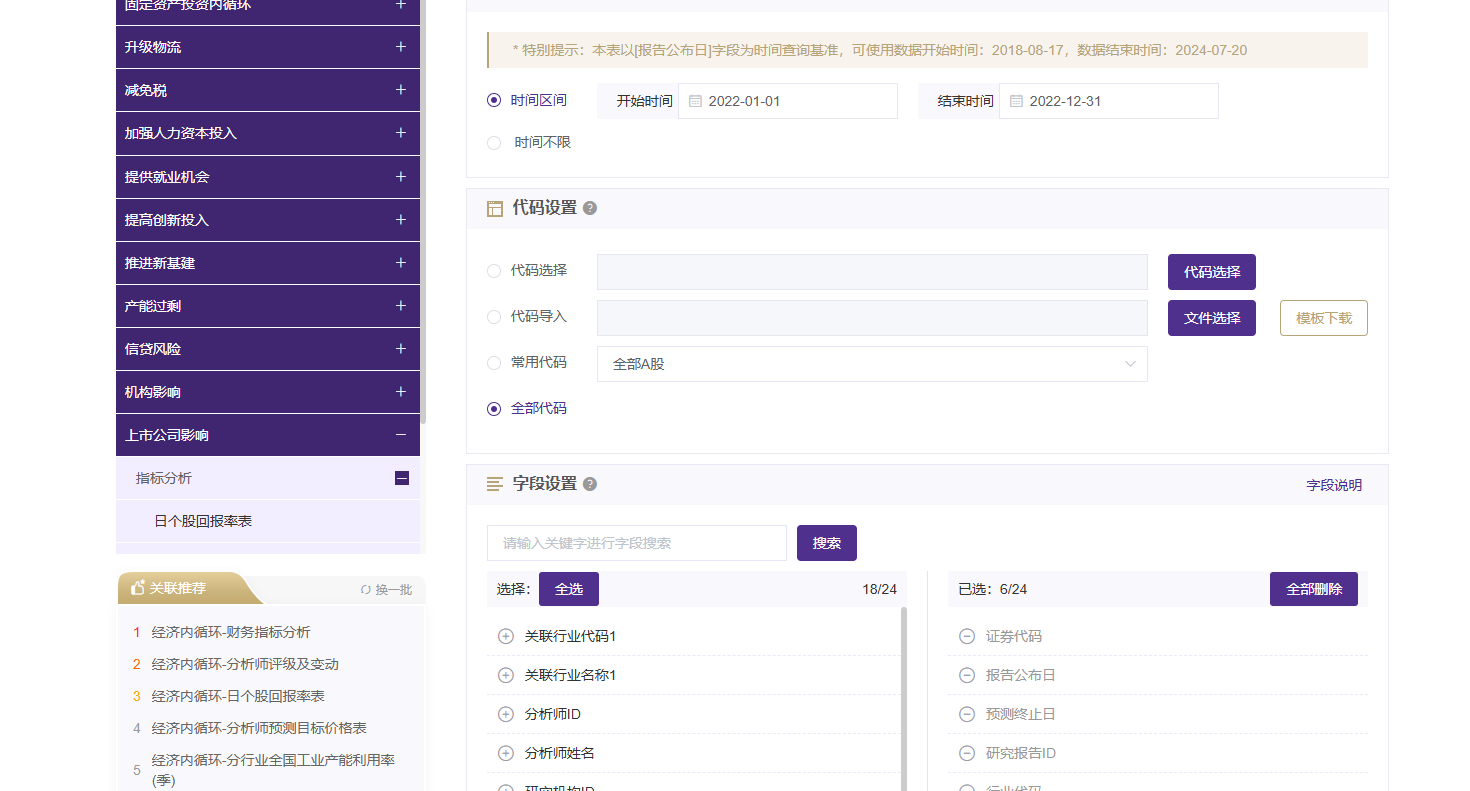
## 2、



使用sql数据进行筛选得到结果并保存，利用python分析，得到2022年的东方财富网的数据总共3283条。

根据此问要求，可能需要用2022-2023的两年的数据（此前理解为一年），但是方法一致。如果需要纯sql获取数据，只需要在此基础上用COUNT方法聚合数据即可。

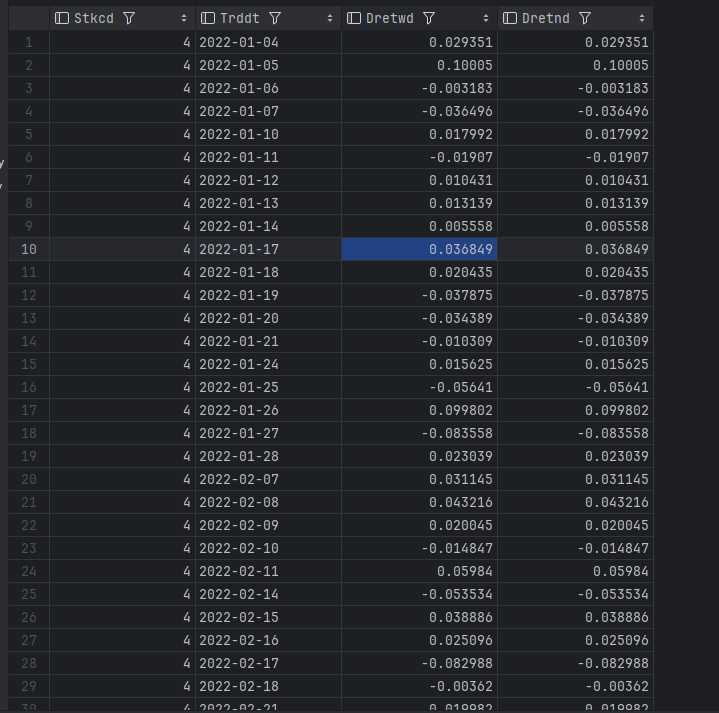
## 3、

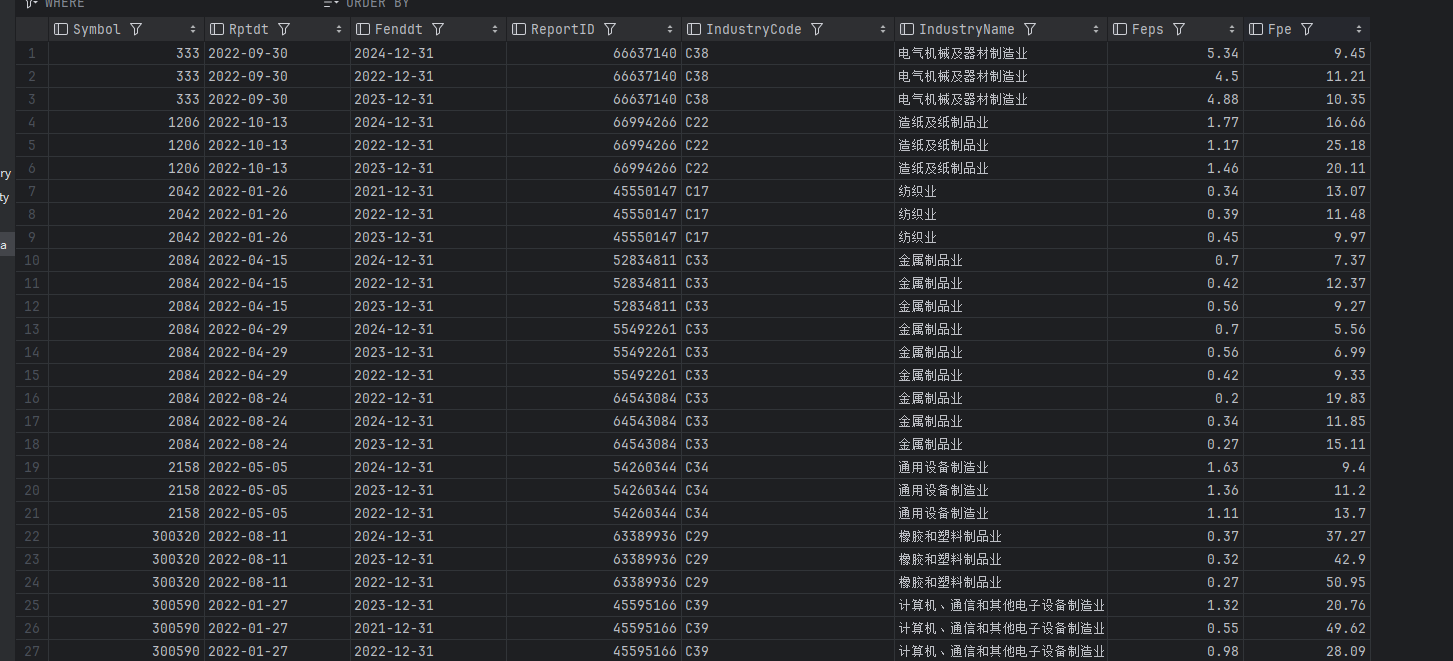


与第一题同理，根据时间筛选数据，并使用python读取，保存至数据库。

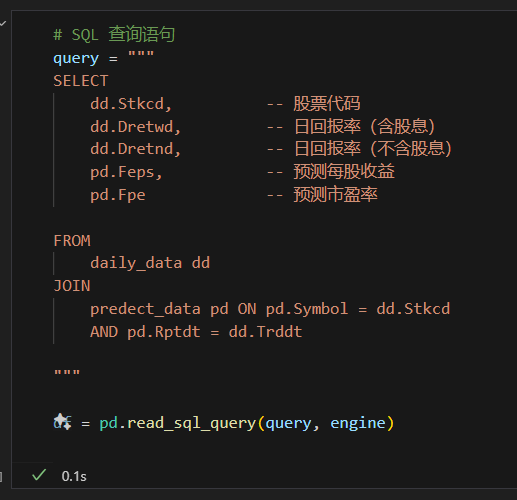
有哪些库 只使用一个库 database

有哪些表 使用两张表 分别为预测信息表和每日数据表

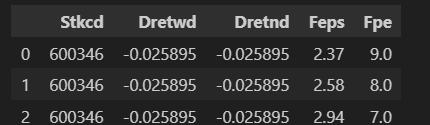




具体的数据处理方式与问题1一致，如下图所示：

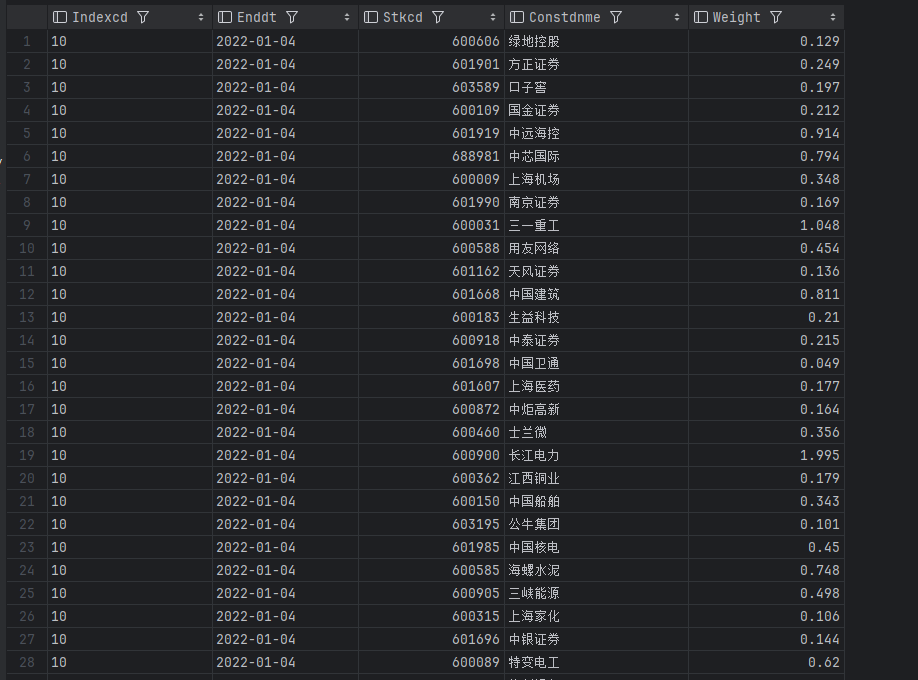


由此得到查询结果：

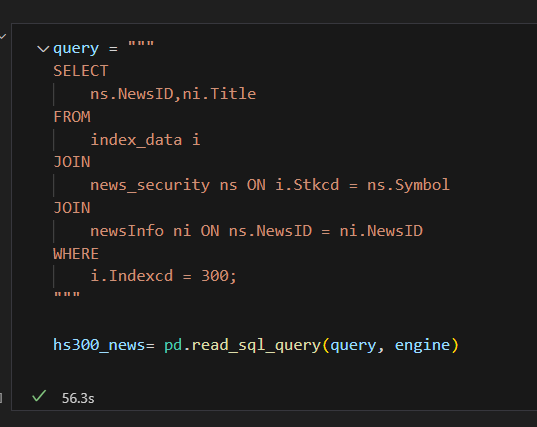


## 4、

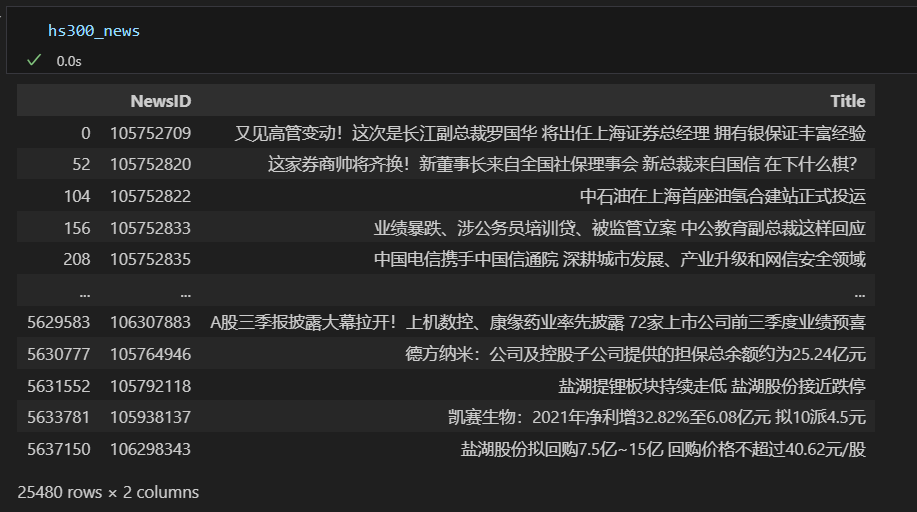
在前三问基础上，还需要获取指数权重表，采用同样方式建表。



使用sql语句获取数据如下：

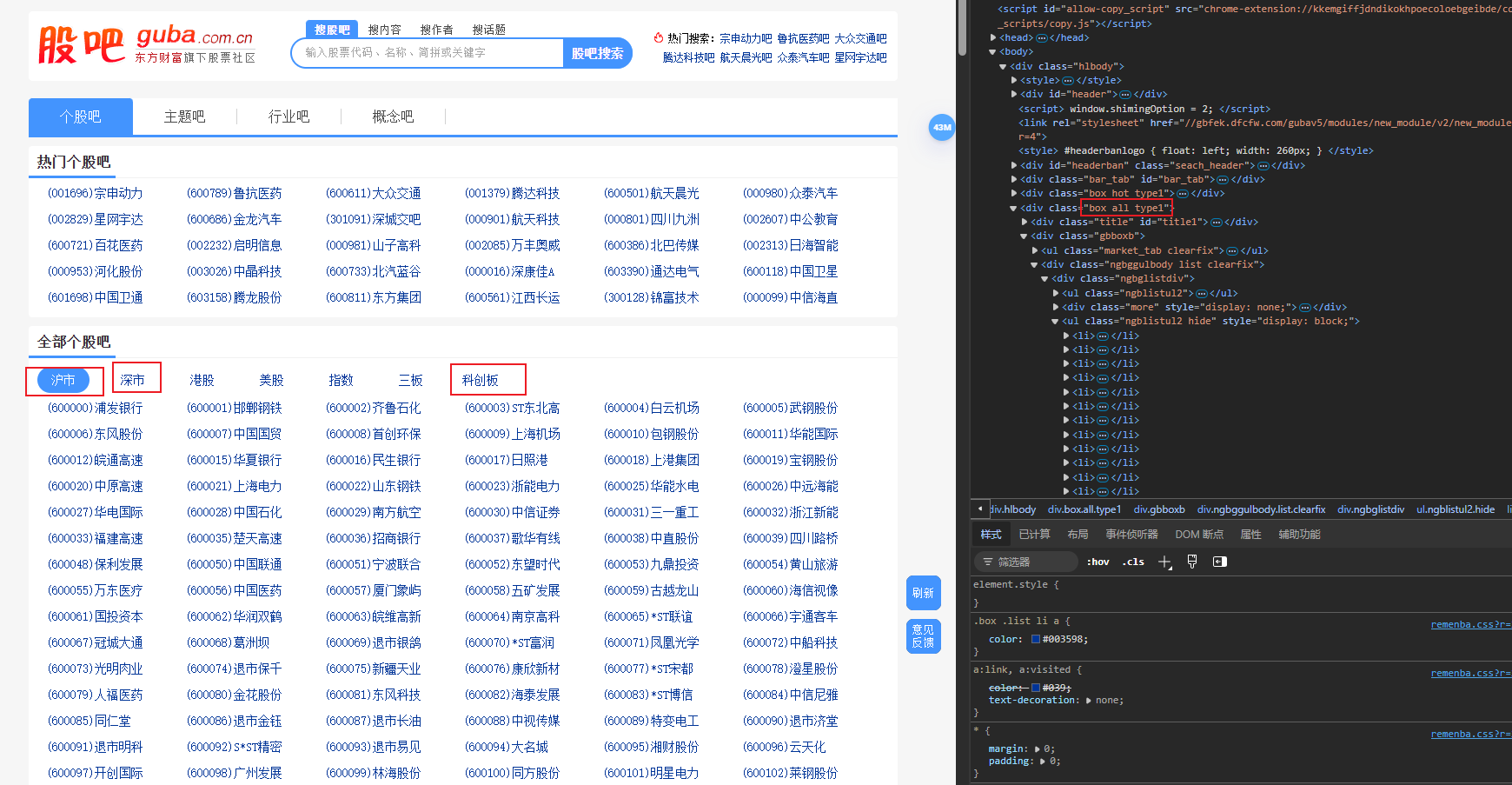


由此得到数据。



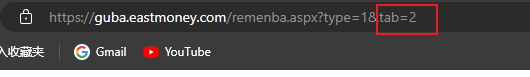
# 二、

选择使用东方财富进行数据爬取。

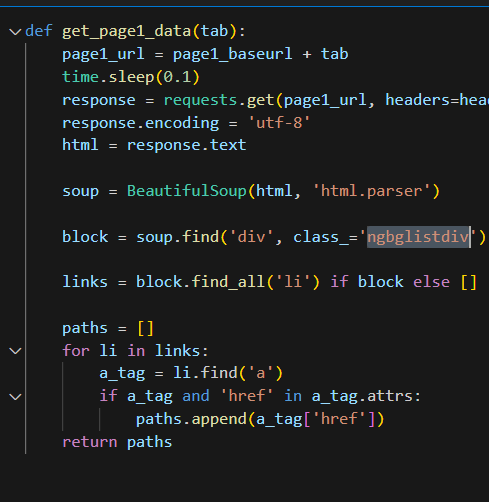


A股数据主要包括沪市、深市和科创版。

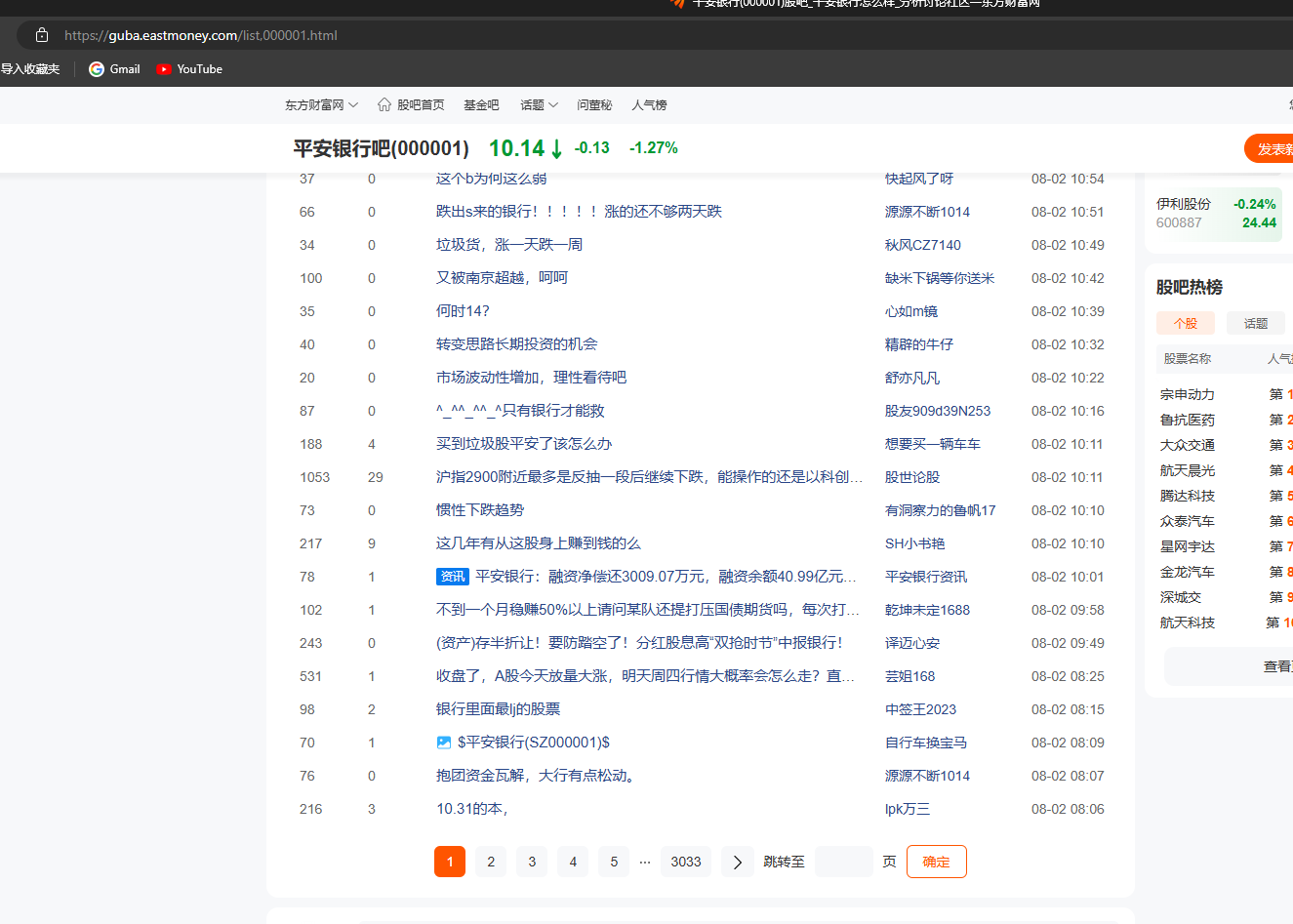
因此进行切换，观察页面，发现不同类型股票由url中的tab控制：



随后解析数据，发现数据在class为ngbglistdiv的所有li元素中，通过beautifulsoup获取每个股票的链接，得到第一个函数：



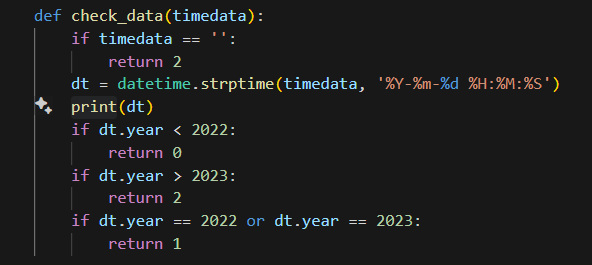
随后进入每个股票的详情链接，如下所示：



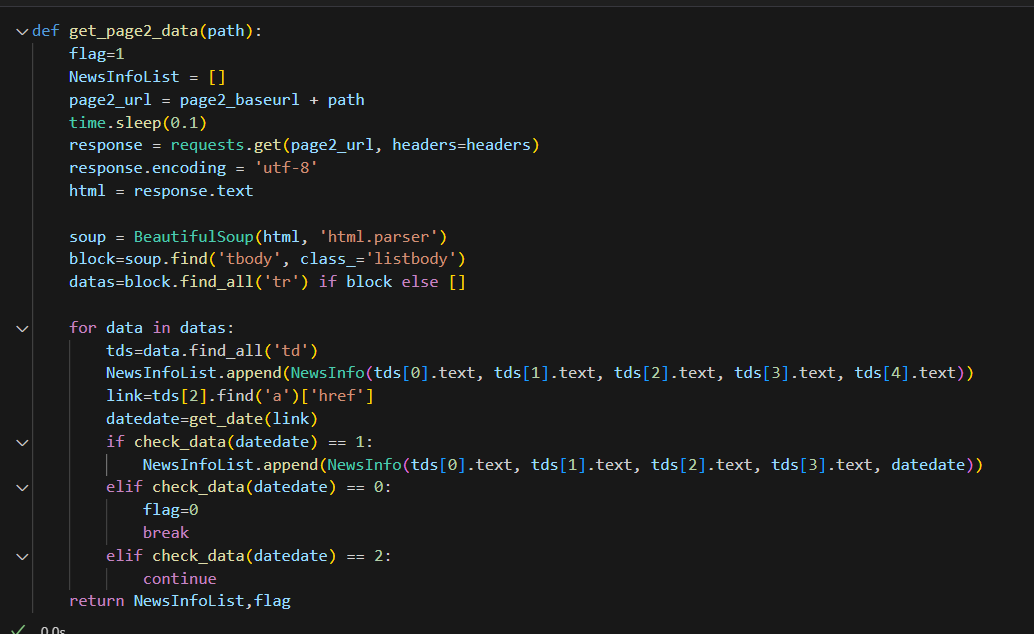
发现数据包括阅读量，评论量、标题、作者和时间，但是不包括年份时间，关于年份时间还需要进一步进入页面爬取，进入下一级页面如下：



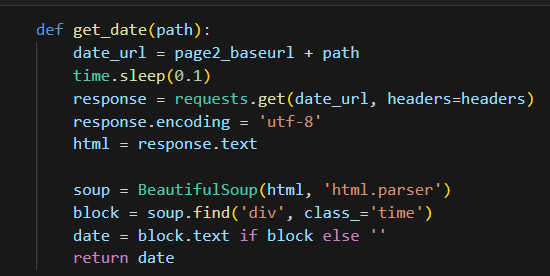
可以得到时间数据，取到时间数据，筛选出2022-2023年的数据，使用函数如下：



并使用一下逻辑处理：



同时包括getdata函数如下：



由此可以爬取到所有数据。

但是在数据爬取过程中，发现网站存在反爬虫机制，高强度的请求会让设备ip在一段时间内被屏蔽导致爬取失败：



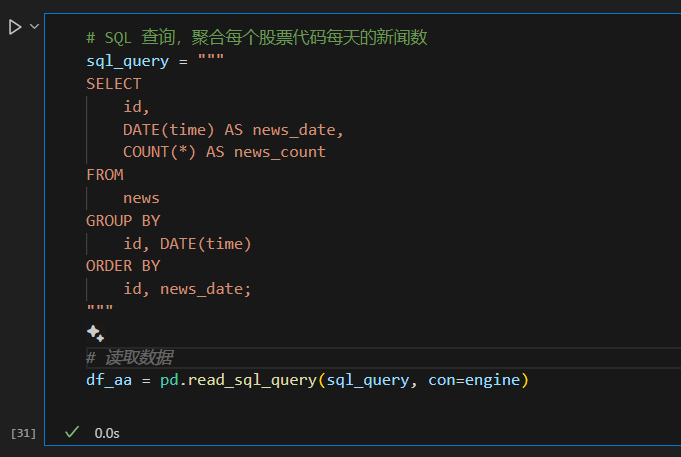
因此只爬取到部分新闻数据，并导入数据库。

数据库设计如下：

使用证券id作为主键，新闻各部分数据作为表数据，建表如下：



由此完成爬虫并存储数据，数据可以通过时间进行日度数据统计处理。

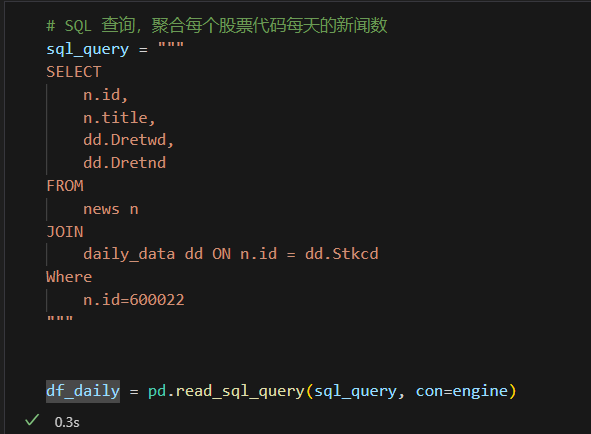


由此可以得到日度数据。

# 三、

在第一和第二题的基础上，查询全部A股的日度收益与你爬取的日度数据。（10分）

只需要根据股票id，使用join进行检索即可。



随机选择一只股票进行检索，得到结果如下：、



可以得到这只股票每日的新闻数据和预测收益。